Published in final edited form as:

Med Image Comput Comput Assist Interv. 2020 October ; 12262: 749–759. doi:10.1007/978-3-030-59713-9_72.

Spatial-Intensity Transform GANs for High Fidelity Medical Image-to-Image Translation

Clinton J. Wang[1], **Natalia S. Rost**[2], **Polina Golland**[1]

[1]Computer Science and Artificial Intelligence Lab, MIT, Cambridge, MA, USA

[2]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA

Abstract

Despite recent progress in image-to-image translation, it remains challenging to apply such techniques to clinical quality medical images. We develop a novel parameterization of conditional generative adversarial networks that achieves high image fidelity when trained to transform MRIs conditioned on a patient's age and disease severity. The spatial-intensity transform generative adversarial network (SIT-GAN) constrains the generator to a smooth spatial transform composed with sparse intensity changes. This technique improves image quality and robustness to artifacts, and generalizes to different scanners. We demonstrate SIT-GAN on a large clinical image dataset of stroke patients, where it captures associations between ventricle expansion and aging, as well as between white matter hyperintensities and stroke severity. Additionally, SIT-GAN provides a disentangled view of the variation in shape and appearance across subjects.

Keywords

Conditional generative adversarial network; Image-to-image translation; Stroke

1 Introduction

Many tasks in medical image analysis require mapping images in one distribution to images in another distribution, conditioned on a set of attributes. Such mappings can be used to synthesize medical images of a specified imaging modality [12] or patient phenotype [9], while preserving most characteristics of an input image such as gross anatomy. Driven by advances in generative adversarial networks (GANs), medical image-to-image translation has been applied to tasks as diverse as data augmentation [1], super-resolution [8], MR-to-CT translation [12], and prediction of disease trajectories [9]. In such GANs, a generator is trained to map input images sampled from a source distribution to synthetic images that appear to belong to a target distribution, while an adversarial discriminator drives the generator to produce realistic images [3,17]. Medical applications of GANs have often been restricted to large datasets of high-quality research scans. When the target distribution is underrepresented in the training data or the data consists of lower quality clinical scans, GANs may introduce severe artifacts.

---

clintonw@csail.mit.edu.

We address this challenge by introducing the spatial-intensity transform generative adversarial network (SIT-GAN), which constrains the generator output to transformations composed of a smooth deformation field and a sparse intensity difference map applied to the input image. This parameterization produces images with fewer artifacts and high fidelity (Fig. 1), and also yields separate visualizations of morphological and tissue intensity changes, which can be relevant to identifying and characterizing disease processes.

Previously, spatial transforms have been coupled with intensity transforms for performing medical image registration [4,16] and data augmentation in the context of semi-supervised segmentation [1]. To the best of our knowledge, this is the first work demonstrating that they are effective at regularizing the outputs of conditional generative models.

Our novel representation of image changes is complementary to prior work on conditional GANs that modify the loss function in the context of simulating aging in brain MRIs. Xia et al. [14] introduce identity-preservation and self-reconstruction losses that penalize large changes in the image for small translations in age. Ravi et al. [9] introduce biological constraints that encourage the network to follow a known hallmark of neurodegeneration, e.g., voxels should darken with age at a rate similar to neighboring voxels. The proposed representation is orthogonal to such changes in the loss function and could be combined to further improve the translation results.

We demonstrate the proposed method on a large dataset of clinical quality brain MRIs of stroke patients. Our experiments suggest that in such settings, SIT-GAN outperforms the state of the art on medical image-to-image translation.

2 Methods

2.1 Image-to-Image Translation with Partially Observed Attributes in Cross-Sectional Data

Given coordinate space Ω, a set $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{N}$ of images $x_i \in \mathscr{X}: \Omega \to \mathbb{R}$ and conditional attributes $y_i \in \mathscr{Y}$ (e.g., age and stroke severity), we want to train a generator to transform images such that their conditional attributes are shifted by a specified amount. Our network consists of a generator $G: \mathscr{X} \times \mathscr{Y} \to \mathscr{X}$, discriminator $D: \mathscr{X} \to \mathbb{R}$ (logits), and regressor $R: \mathscr{X} \to \mathscr{Y}$. Here we consider continuous vector attributes $y_i = (y_{i,1}, \ldots, y_{i,m})$ that may have missing values. Categorical attributes can be included by expanding regressor R to produce categorical outputs (classifier).

Generator.—The generator G transforms an input image such that the transformed image appears to take on different attribute values from the input image, but maintains aspects of the input image that are unrelated to the conditional attributes, such as non-pathological anatomy.

Define $z_i = (x_i, y_i)$, $z_j = (x_j, y_j)$, $\Delta y = y_j - y_i$. During training, the generator is updated using the following loss terms (Fig. 2):

$$\ell_{\mathrm{cc}} = \| G(G(x_i, \Delta y), -\Delta y) - x_i \|_1 \quad \text{cycle consistency loss} \tag{1}$$

$$\ell_{\text{attr}} = \frac{1}{m} \| (R(G(x_i, \Delta y)) - R(x_i)) - \Delta y \|_2^2 \quad \text{relative attribute loss} \tag{2}$$

$$\ell_{\text{adv}} = -D(G(x_i, \Delta y)) \quad \text{Wasserstein adversarial loss} \tag{3}$$

Parameterizing generator G in terms of attribute difference Δy enables evaluation of the cycle consistency loss ℓ_{cc} even when images have missing attributes [13]. To compute Δy in such cases, we introduce the convention that $y_{j,k}-y_{i,k} = 0$ if either attribute is missing. Putting the terms together,

$$\mathscr{L}_G = \mathbb{E}_{z_i, z_j}[\ell_{\text{adv}} + \lambda_{\text{attr}}\ell_{\text{attr}} + \lambda_{\text{cc}}\ell_{\text{cc}}] \tag{4}$$

where λ_{attr} and λ_{cc} are empirically determined weights.

Discriminator.—We simultaneously train the discriminator D with the Wasserstein GAN losses and gradient penalty [6]:

$$\mathscr{L}_D = \mathbb{E}_{z_i, z_j}[D(G(x_i, \Delta y))] - \mathbb{E}_{z_i}[D(x_i)] - \mathbb{E}_{\hat{x}}\left[\lambda_{GP}(\| \nabla_{\hat{x}} D(\hat{x}) \|_2 - 1)^2\right] \tag{5}$$

where $\hat{x}$ is obtained by interpolating real and translated images as described in [6], and λ_{GP} is a weight.

Regressor.—The regressor R is trained to predict the attributes of real images, using a mean squared error loss.

$$\mathscr{L}_R = \mathbb{E}_{z_i}\left[\frac{1}{m} \| R(x_i) - y_i \|_2^2\right] \tag{6}$$

We share layers between the discriminator and regressor, so a single optimizer is assigned to both subnetworks and updated using $\mathscr{L}_D + \lambda_R \mathscr{L}_R$.

2.2 Spatial-Intensity Transform Generator

To constrain the generator to spatial-intensity transforms, we define its outputs as the deformation field $F:\Omega \to \mathbb{R}^d$ for image dimensionality d, with corresponding transform $T_F:\mathscr{X} \to \mathscr{X}$,, and the intensity difference map $\Delta x:\Omega \to \mathbb{R}$..

Rather than directly producing the target image, the generator outputs the deformation field F and intensity changes Δx, then transforms the input image as $T_F(x_{\text{in}} + \Delta x)$. In addition, we added regularization terms to the generator's loss function that encourage the deformation field to be smooth and the intensity difference map to be sparse. Specifically, we used the discrete total variation norm [2] to regularize the deformation field and the L1-norm to regularize the intensity change:

$$\| F \|_{\mathrm{TV}} = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \| \nabla F(\omega) \|_2 \tag{7}$$

$$\| \Delta x \|_1 = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} |\Delta x(\omega)|, \tag{8}$$

where $\|\nabla F(\omega)\|$ is approximated using finite differences. The total generator loss becomes $\mathscr{L}_G = \mathbb{E}_{z_i, z_j}[\ell_{\mathrm{adv}} + \lambda_{\mathrm{attr}}\ell_{\mathrm{attr}} + \lambda_{\mathrm{cc}}\ell_{\mathrm{cc}} + \lambda_{\mathrm{TV}} \| F \|_{\mathrm{TV}} + \lambda_{\Delta x} \| \Delta x \|_1]$ for empirically determined weights λ_{TV} and $\lambda_{\Delta x}$.

2.3 Network Architecture and Implementation Details

SIT-GAN's generator was implemented as a 2D U-Net that takes in attribute difference Δy by replicating each dimension of Δy spatially and concatenating channel-wise with the input image x_{in}. The U-Net has 4 spatial resolutions, with 200 channels and 6 residual blocks at the lowest resolution. The discriminator and regressor share 5 down-sampling blocks, then split into fully connected layers of the appropriate dimension (1 output for the discriminator, m outputs for the regressor).

Batch normalization is used for all convolutional layers. Down-sampling blocks in the U-Net use convolutional layers alternating with max blur pooling [15]. Up-sampling blocks in the U-Net use bilinear upsampling between convolutional layers. The generator uses ReLU activations, the discriminator and regressor use leaky ReLU activations.

The subnetworks were trained with Adam optimizers, with one step in G's optimizer for every two steps in D/R's optimizer. D/R were trained for 50K iterations with a learning rate of 1.2×10^{-5}, and G was trained for 25K iterations with a learning rate of 1.5×10^{-4}. Both optimizers used a minibatch size of 4, and moving average parameter $\beta_1 = 0.86$. We used the following loss weights: $\lambda_R = 18$, $\lambda_{\mathrm{attr}} = 3.5$, $\lambda_{\mathrm{cc}} = 2.1$, $\lambda_{\mathrm{TV}} = 16$, $\lambda_{\Delta x} = 49$, and $\lambda_{\mathrm{GP}} = 1.1$.

2.4 Baseline Networks

To investigate the influence of different components, we compare SIT-GAN to a network whose generator does not transform the image and several networks whose generators use alternate transformations of the input image. These different parameterizations are summarized in Table 1.

In the unconstrained network, the generator directly synthesizes a new image. We trained two variants of this model: one that has identical hyperparameters to SIT-GAN and other baseline networks, and one in which we tuned the number of layers, types of layers, loss term weights, and type of optimizer to make it as competitive with SIT-GAN as possible. In the tuned model, the discriminator and regressor were trained with a learning rate of 8.6×10^{-5}, and the generator was trained with a learning rate of 1.1×10^{-4}. The U-Net had 3 spatial resolutions with 96 channels and 3 residual blocks at the lowest resolution. Strided convolutions were used for downsampling. The discriminator/regressor had 6 downsampling blocks using max blur pooling. The tuned network used loss weights of $\lambda_R = 21$, $\lambda_{\mathrm{attr}} = 1$, $\lambda_{\mathrm{cc}} = 4$, and $\lambda_{\mathrm{GP}} = 8$. The Adam optimizer had moving average parameter $\beta_1 = 0.46$.

In the difference transform network, the generator is constrained to a sparse intensity difference transform of the input image. It penalizes output images that differ from their inputs using the L1-norm, making it suitable for capturing image-to-image translations that only involve small regions of the image. It corresponds to SIT-GAN with $\lambda_{TV} = \infty$.

The optical flow network is constrained to smooth deformations of the input image, which can capture morphological variation but not intensity changes within anatomical structures. It corresponds to SIT-GAN with $\lambda_{\Delta x} = \infty$.

The weighted flow network outputs a weighted sum of the input image and a smooth deformation of it, with pixel-wise weights computed by the generator. It is the type of model used to synthesize successive frames in video-to-video translation models [11].

2.5 Data and Evaluation

Our dataset consisted of 1821 axial brain fluid-attenuated inversion recovery (FLAIR) MRIs from 12 clinical sites in the MRI-GENIE study [5], obtained within 48 h of symptom onset in acute ischemic stroke patients. 418 images acquired from the largest site (Massachusetts General Hospital) were used for 5-fold cross validation. The models were then tested on the 1403 scans from all other clinical sites. Age was available for all patients, and stroke severity (NIHSS scale of 0–36) was available for 746 patients.

MRIs were preprocessed with resampling to isotropic 1mm resolution, N4 bias field correction, ANTS registration to a FLAIR atlas, normalisation of the white matter intensity, and cropping to 224 × 192. Native resolution varied, but was typically around 1mm × 1mm × 6mm, which resulted in significant partial volume effects. The 15 middle axial slices of each subject were used, and all slices from the same subject were grouped into the same validation fold. We scaled age and stroke severity so that the empirical distribution of each attribute within the training data had a mean of 0 and a standard deviation of 1. The images were also augmented using horizontal flips and random affine transformations.

To quantify the realism of model outputs in the absence of paired data, we computed the Fréchet Inception Distance (FID) [7] between the distribution of generated images and the distribution of validation or test images. We also used Precision and Recall for Distributions (PRD) [10] to compute the $F_{1/8}$ and F_8 scores of our generator. A high $F_{1/8}$ suggests that most modes of the generated distribution belong to the true distribution, whereas a high F_8 suggests that most modes of the true distribution belong to the generated distribution. Modes are estimated by finding clusters of images in Inception v3 embedding space.

We also evaluated the effectiveness of each model in transforming the target attribute by measuring the performance of an Inception v3 regressor on our generated images. This regressor was pre-trained on ImageNet and fine-tuned on FLAIR MRIs to predict both age and stroke severity. We emphasize that this Inception v3 regressor is different from the regressor used during training of the GAN, as the generator may have learned to exploit peculiarities in the particular regressor it is trained with. Using a separately trained regressor with a different architecture eliminates any gains that the generator accrued in this manner. We measure the mean squared error (MSE) of age and stroke severity (NIHSS) respectively, normalized to the empirical standard deviation of the attribute. The MSE of the Inception regressor on held out subjects in the cross-validation set is 0.24 on age and 0.70 on NIHSS, while it is 0.34 on age and 0.62 on NIHSS in the test set.

3 Results

Our results suggest that SIT-GAN achieves better image fidelity than the unconstrained model as measured by FID, precision ($F_{1/8}$) and recall (F_8). The statistically significant improvement in both cross-validation and testing (p<0.01 for each metric by t-test) shows that this pattern generalizes beyond the particular clinical site it was trained on.

Even after tuning, the unconstrained model introduces artifacts in translated images such as dark streaking of the gray matter with increasing age, and partial volume-like filling of the ventricles with decreasing age. While some of these artifacts do appear in the dataset, Fig. 3 illustrates such artifacts in images that did not have them originally. SIT-GAN does not suffer from these artifacts, and still captures the growth of the ventricles correlated with aging.

The difference transform, optical flow, weighted flow, and SIT-GAN models all perform relatively well on distributional metrics but underperform the unconstrained model on target domain transfer (Table 2). This suggests that an unconstrained generator sacrifices image quality to capture more variation in the conditional attribute compared to the constrained generators.

In general, SIT-GAN attains the best image fidelity, and performs similarly to the optical and weighted flow models in target distribution matching. Often it is overly conservative in transforming input images, but when it succeeds, it is able to capture the expansion of the ventricles correlated with aging as well as the increase in white matter hyperintensities associated with stroke severity (Fig. 1), while producing less severe artifacts than the unconstrained model as seen in Fig. 3.

The learned spatial and intensity transforms also correspond to changes in morphology and tissue properties that are associated with particular patient phenotypes or disease processes. In FLAIR MRIs of stroke patients, the patient’s age is correlated with the volume of the ventricles as well as the volume of white matter hyperintensities in the periphery of the ventricles. These effects, which would be inseparable in the unconstrained model, can be visualized separately with SIT-GAN by examining the deformation field and intensity differences individually (Fig. 4).

4 Conclusion

We presented SIT-GAN, a novel parameterization of GANs for medical image-to-image translation that improves image fidelity and reduces artifacts. In many medical applications, the desired transformations can be well represented by a smooth deformation and a sparse intensity difference transform, and our method can provide robustness to artifacts. We demonstrated our model on a challenging dataset of clinical quality FLAIR MRIs of stroke patients. Our model produces high quality images that visualize correlations of the brain’s shape and appearance with the patient's age and stroke severity. Additionally, our parameterization provides a disentangled view of changes in anatomical shape and tissue appearance. Such advances in image-to-image translation can help drive progress in many areas of medical image analysis, including data augmentation, data harmonization, and prediction or visualization of disease trajectories.

Because our proposed representation sacrifices the quality of target domain matching to improve image fidelity, our work leaves open questions about how to navigate or circumvent this trade-off. We suggest that promising directions include carefully relaxing the constraints (reducing regularization weights) over the course of training, or incorporating priors over anatomical structures. Future work will also extend our 2D model to 3D, using low-memory techniques to compensate for memory limitations, and more aggressive data augmentation to accommodate higher model capacity.

Acknowledgments.

This work was supported by NIH NIBIB NAC P41EB015902, NIH NINDS U19NS115388, NIH NINDS R01NS086905, Wistron Corporation, Takeda Pharmaceuticals and the Siebel Foundation.

References

1. Chaitanya K, Karani N, Baumgartner CF, Donati O, Becker AS, Konukoglu E: Semi-supervised and task-driven data augmentation. CoRR abs/1902.05396 (2019). http://arxiv.org/abs/1902.05396
2. Chambolle A, Novaga M, Cremers D, Pock T: An introduction to total variation for image analysis. In: Theoretical Foundations and Numerical Methods for Sparse Recovery, De Gruyter (2010)
3. Choi Y, Choi M, Kim M, Ha J, Kim S, Choo J: StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. CoRR abs/1711.09020 (2017). http://arxiv.org/abs/1711.09020
4. Cootes TF, Beeston C, Edwards GJ, Taylor CJ: A unified framework for atlas matching using active appearance models In: Kuba A, Šáamal M, Todd-Pokropek A (eds.) Information Processing in Medical Imaging, pp. 322–333. Springer, Heidelberg (1999)
5. Giese A, et al.: Design and rationale for examining neuroimaging genetics in ischemic stroke: the MRI-genie study. Neurol. Genet. 3(5) (2017). 10.1212/NXG.0000000000000180
6. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC: Improved training of wasserstein gans. CoRR abs/1704.00028 (2017). http://arxiv.org/abs/1704.00028
7. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Klambauer G, Hochreiter S: Gans trained by a two time-scale update rule converge to a nash equilibrium. CoRR abs/1706.08500 (2017). http://arxiv.org/abs/1706.08500
8. Quan TM, Nguyen-Duc T, Jeong WK: Compressed sensing MRI reconstruction using a generative adversarial network with a cyclic loss. IEEE Trans. Med. Imaging 37(6), 1488–1497 (2018) [PubMed: 29870376]
9. Ravi D, Alexander DC, Oxtoby NP: Degenerative adversarial neuroimage nets: generating images that mimic disease progression In: Shen D, et al. (eds.) MICCAI 2019. LNCS, vol. 11766, pp. 164–172. Springer, Cham (2019). 10.1007/978-3-030-32248-9_19
10. Sajjadi MSM, Bachem O, Lucic M, Bousquet O, Gelly S: Assessing generative models via precision and recall (2018). https://arxiv.org/abs/1806.00035
11. Wang T, et al.: Video-to-video synthesis. CoRR abs/1808.06601 (2018). http://arxiv.org/abs/1808.06601
12. Wolterink JM, Dinkla AM, Savenije MHF, Seevinck PR, van den Berg CAT, Išgum I: Deep MR to CT synthesis using unpaired data In: Tsaftaris SA, Gooya A, Frangi AF, Prince JL (eds.) SASHIMI 2017. LNCS, vol. 10557, pp. 14–23. Springer, Cham (2017). 10.1007/978-3-319-68127-6_2

13. Wu PW, Lin YJ, Chang CH, Chang EY, Liao SW: RelGAN: multidomain image-to-image translation via relative attributes. In: International Conference on Computer Vision (2019)
14. Xia T, Chartsias A, Wang C, Tsaftaris SA: Learning to synthesise the ageing brain without longitudinal data (2019). https://arxiv.org/abs/1912.02620
15. Zhang R: Making convolutional networks shift-invariant again. CoRR abs/1904.11486 (2019). http://arxiv.org/abs/1904.11486
16. Zhao A, Balakrishnan G, Durand F, Guttag JV, Dalca AV: Data augmentation using learned transforms for one-shot medical image segmentation. CoRR abs/1902.09383 (2019). http://arxiv.org/abs/1902.09383
17. Zhu J, Park T, Isola P, Efros AA: Unpaired image-to-image translation using cycle-consistent adversarial networks. CoRR abs/1703.10593 (2017). http://arxiv.org/abs/1703.10593

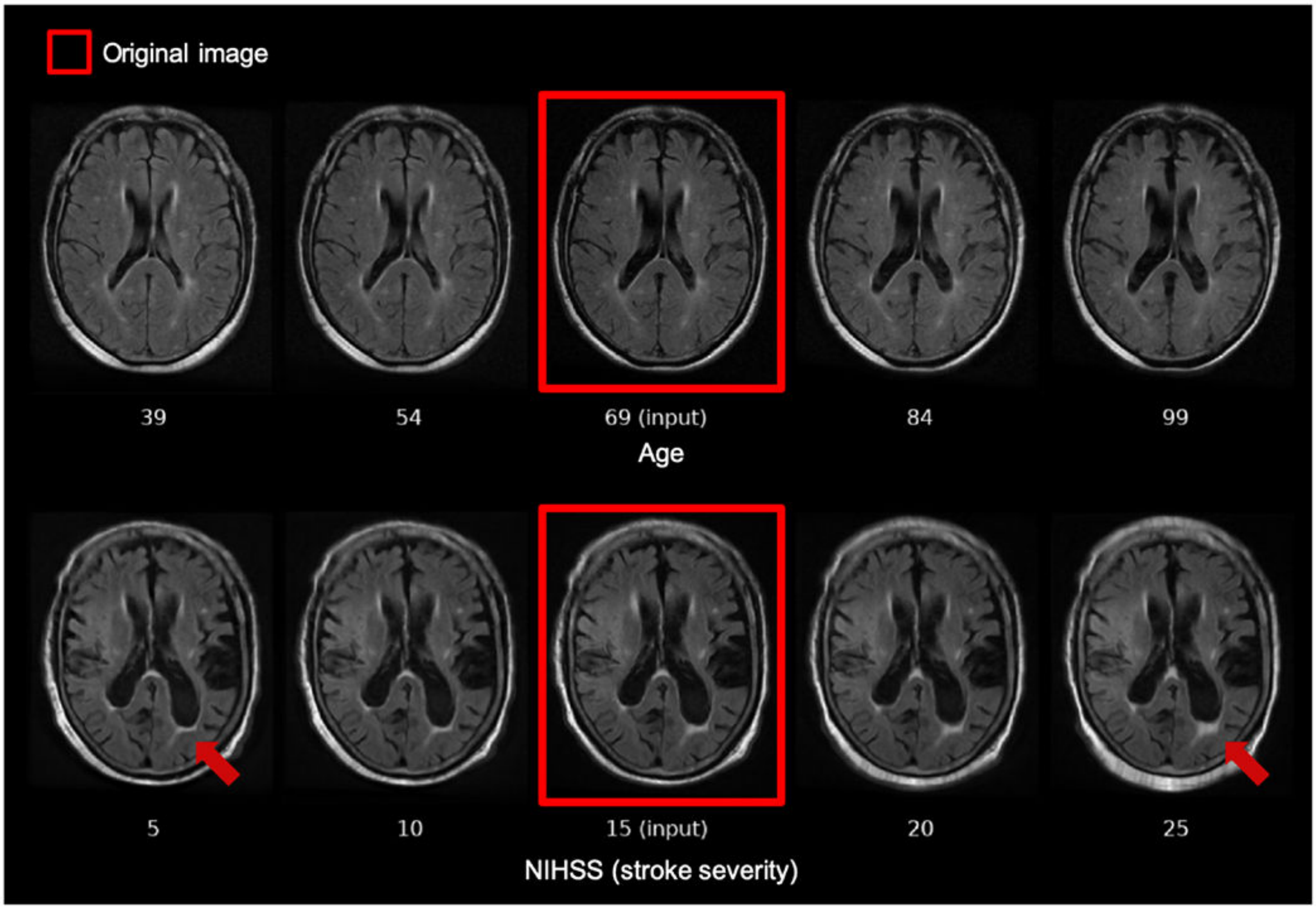

Fig. 1.

Synthetic fluid-attenuated inversion recovery (FLAIR) MRIs of acute ischemic stroke patients, obtained by transforming an input MRI (center column) conditioned on changes in age (top) and stroke severity (bottom). Increasing age correlates with increasing ventricular volume, and increasing stroke severity correlates with increasing volume of periventricular white matter hyperintensities (see red arrow).

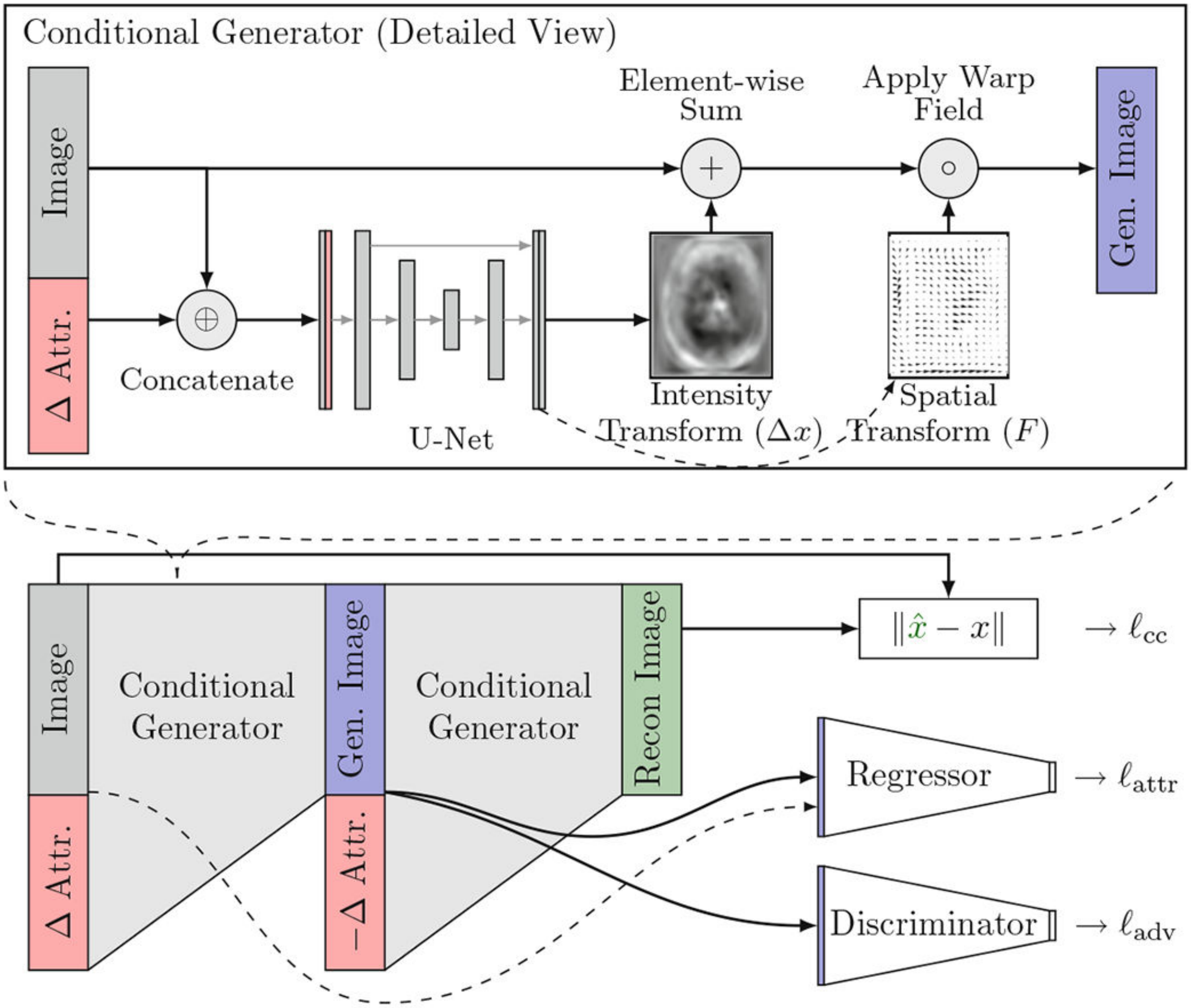

Fig. 2.

The generator takes in an image and the desired change in each attribute. In our spatial-intensity transform GAN, the generated image is obtained by applying an intensity difference map and a deformation field to the input image. The parameters of the generator are updated from three loss terms: a cycle consistency loss ℓ_{cc} that discourages unnecessary changes to the input image, an attribute loss ℓ_{attr} that encourages the generated image to match the desired attribute values, and an adversarial loss ℓ_{adv} that penalizes unrealistic generated images.

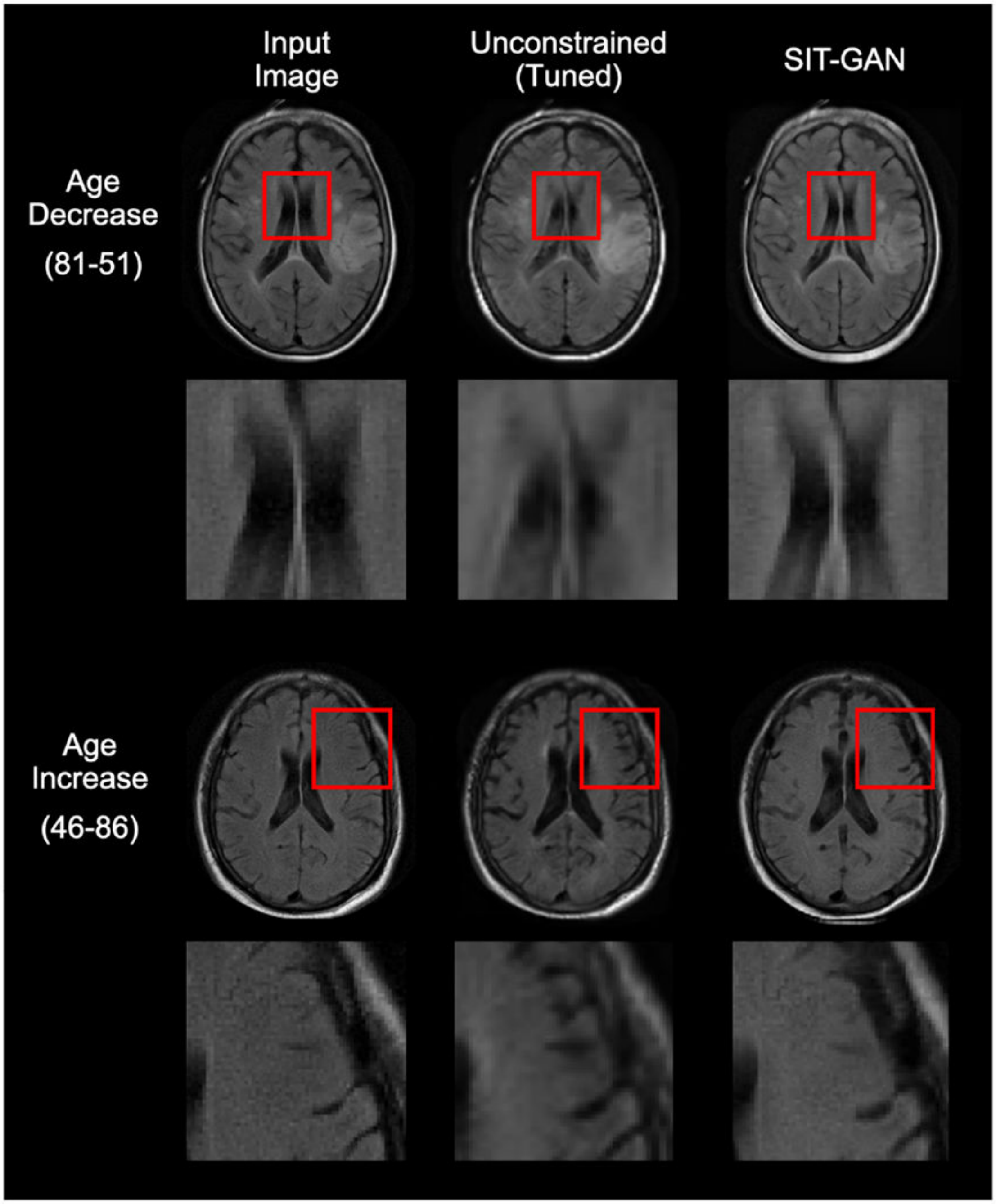

Fig. 3.

Comparison of stroke MRIs translated to a different age using the baseline model and our model. While both models change the ventricle shape appropriately, the baseline model blurs the ventricles (top rows) and excessively darkens the gray matter (bottom rows).

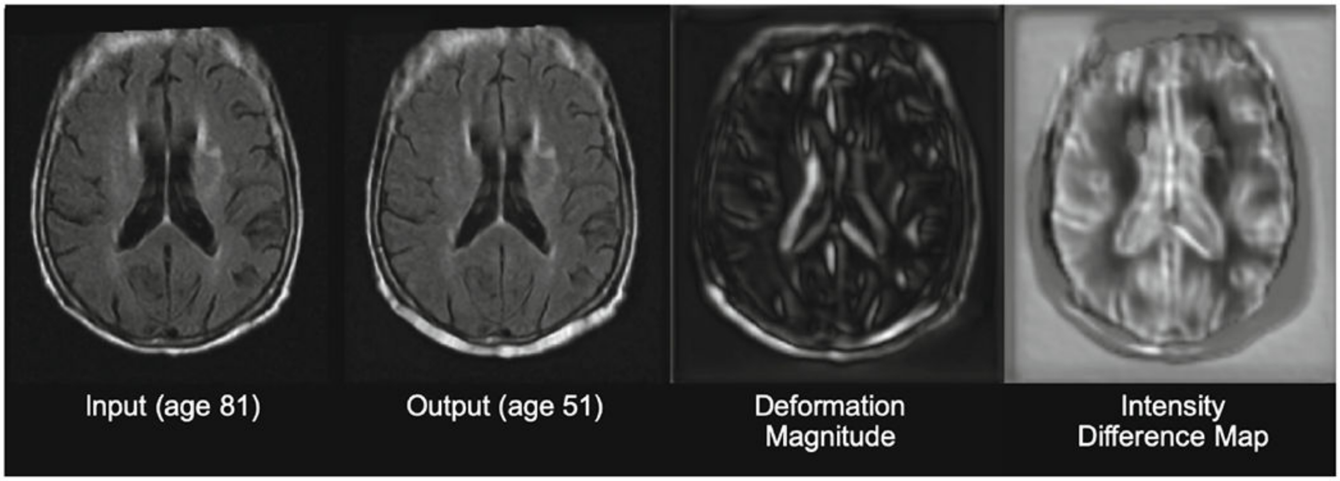

Fig. 4.

The magnitude of the deformation field and intensity difference map of the spatial-intensity model for an example transformation. The shrinkage of the ventricles and sulci are well captured by the deformation field, while tissue appearance changes are reflected in the difference map.

Table 1.

Parameterizations of the generator output.

Parameterization	G Outputs	Generated image	Regularizers
Unconstrained	x_{out}	x_{out}	N/A
Difference Transform	Δx	$x_{\text{in}} + \Delta x$	$\|\Delta x\|_1$
Optical Flow	F	$T_F(x_{\text{in}})$	$\|F\|_{\text{TV}}$
Weighted Flow	F, w	$w \odot T_F(x_{\text{in}}) + (1 - w) \odot x_{\text{in}}$	$\|F\|_{\text{TV}}$
SIT-GAN	$F, \Delta x$	$T_F(x_{\text{in}} + \Delta x)$	$\|F\|_{\text{TV}}$, $\|\Delta x\|_1$

Table 2.

Performance metrics for translation of FLAIR MRIs conditioned on age and stroke severity (NIHSS), averaged over 5 runs. FID = Fréchet Inception Distance, P/R = Precision ($F_{1/8}$) and Recall (F_8) as defined in [10].

Model Type	**FID**	**P/R**	**Age MSE**	**NIHSS MSE**
Cross-validation				
Unconstrained	152.1	0.01/0.01	1.51	2.18
Unconstrained (tuned)	61.4	0.07/0.21	**0.51**	1.12
Difference transform	57.2	**0.38**/**0.59**	1.37	1.14
Optical flow	59.5	0.30/0.52	0.71	**1.09**
Weighted flow	60.6	0.23/0.46	0.85	1.31
SIT-GAN	**38.6**	0.35/**0.59**	0.85	1.16
Test				
Unconstrained	180.5	0.07/0.02	1.11	1.21
Unconstrained (tuned)	51.0	0.41/0.21	**0.99**	**1.01**
Difference transform	68.4	0.53/0.68	1.25	1.12
Optical flow	28.4	**0.62**/**0.69**	1.16	1.11
Weighted flow	35.0	0.56/0.59	1.32	1.14
SIT-GAN	**27.6**	0.53/0.66	1.28	1.12